I.J. Modern Education and Computer Science, 2015, 2, 42-47
Published Online February 2015 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2015.02.06

Improving the Efficiency of Term Weighting in Set of Dynamic Documents

Mehdi Jabalameli
Department of Computer Engineering, University of Isfahan, Isfahan, 8174673441, Iran
Email: jabalameli@eng.ui.ac.ir

Ala Arman
KTH Royal Institute of Technology, Stockholm, SE-100 44, Sweden
Email: aarman@kth.se

Mohammadali Nematbakhsh
Department of Computer Engineering, University of Isfahan, Isfahan, 8174673441, Iran
Email: nematbakhsh@eng.ui.ac.ir

Abstract—In real information systems, there are few static documents. On the other hand, there are too many documents that their content change during the time that could be considered as signals to improve the quality of information retrieval. Unfortunately, considering all these changes could be time-consuming. In this paper, a method has been proposed that the time of analyzing these changes could be reduced significantly. The main idea of this method is choosing a special part of changes that do not make effective changes in the quality of information retrieval; but it could be possible to reduce the analyzing time. To evaluate the proposed method, three different datasets selected from Wikipedia. Different factors have been assessed in term weighting and the effect of the proposed method investigated on these factors. The results of empirical experiments showed that the proposed method could keep the quality of retrieved information in an acceptable rate and reduce the documents' analysis time as a result.

Index Terms—Document Revision, Term Frequency, Term Weightings, Ranked Terms, Information retrieval process.

I. Introduction

Nowadays, web pages are dynamic and during time their information is changed. In many applications such as search engines, each change in one page is considered as a new page. However, these pages are actually the different versions of the same page. For example, the Wikipedia pages have different versions that these versions are created by different people to improve the content of the pages. Previous researches [1-11] have shown that investigating on these changes can improve the efficiency of information retrieval systems.

In information retrieval systems, usually a document is seen as a vector of terms which each component of the vector shows the one of the terms' weight. Reference [12] provides the popular formula of term weighting is called TF_IDF which the number of repetitive terms and the number of all documents on a set that include that term are used for term weighting. In the most of methods that have been proposed for term weighting, a document has been considered statically. However, the content of many of these pages change during the time and these changes could be containing important information in the retrieval process.

In the recent years, some researches have been done that changes in a document (or a page) have been considered in the quality of information retrieval and term weighting. In the second section of this paper, some these researches will be reviewed.

In all of these researches, the whole document and its previous histories should be analyzed to specify the weight of each term and the relatedness of extracting documents with the query. In this paper, for the first time, the run time of these algorithms has been considered. In the proposed method, instead of analyzing the whole history of documents, only a special part of documents' histories is analyzed and the weight of terms is specified.

Empirical experiments have been done on the three different datasets of Wikipedia. The results have been shown that the quality of term weighting in the proposed method is almost the same as existing ones. However, the processing time of the investigation of documents' records have been significantly reduced due to selecting a special part of documents' histories.

The rest of paper is organized as follows. Section II, provides some related works in term weighting using different methods. Section III, presents the proposed method to improve the term weighting efficiency. In section IV, the empirical evaluation of the suggested method is described. Finally, conclusion and future works are discussed in section V.

II. Related Works

A. Global Term Weighting in a Set of Documents

Reference [6], considers the dynamicity of documents for the first time and the weight of terms has been defined based on the level of documents' dynamicity. Empirical experiments in this research have shown that the pages that have the most relation with the user's queries change more than other pages. For this reason, in this paper, these changes have been used for term weighting. To do this, the total time of changes in a document has been broken to some specified intervals (T=10) and the frequency of terms has been used in each interval in term weighting. This weighting has been used for document ranking.

In general, the ranking of document D for query Q is calculated by

$$P(D|Q)=P(D)P(Q|D)=P(D)\prod_{q\in Q} P(q|D)^{n(q,Q)} \quad (1)$$

where, D is a document in a special time slice, q ∈Q is a term in query Q and n(q,Q) is the frequency of the term q in the query Q. Reference [6], considers the frequency of the q in the dynamic document D has been considered as the following vector in (2):

$$n(q, D) = < n(q, D^1), n(q, D^2), \dots, n(q, D^T) > \quad (2)$$

where, n (q, D^I) shows the frequency of the term q in document D on the time interval T.

To include the changes of the documents in term weighting, the terms of each document have been divided into three groups: long-term, mid-term, and short-term. If c(q,D) is the number of non-zero elements of the vector n(q,D), the membership of each term in each of the groups is based on the value of c (q, D) so that the membership of c (q,D) in each interval of [0.9∗ T, T], [0.5∗ T, 0.9∗T), and [0, 0.5∗T) show short-term, mid-term, and long-term groups respectively.

According to the above grouping, the following probabilistic relationship for the probability of occurrence of q in document D has been suggested as

$$P(q \mid D)=\lambda_L P(q \mid D_L) + \lambda_M P(q \mid D_M) + \lambda_S P(q \mid D_S) \quad (3)$$

where D_S, D_M, and D_L show the documents of short-term, mid-term, and long-term group respectively. Moreover, λ_L, λ_M, and λ_S are configurable parameters so that

$$\lambda_L + \lambda_M + \lambda_S = 1 \quad (4)$$

Empirical experiments have shown that the ranking of documents based on the offering formula, has had positive effects on the quality of document retrieval.

B. Term Weighting Using Time Series

Reference [5], uses time series for term weighting. If the frequency of each term *x* at time *t* in the whole set of documents is shown by x_i, the frequency of *x* at the times t_1, t_2,t_3,…, t_n will make the time series $\{x_1,x_2,x_3,\dots,x_n\}$. In linear time series, each x_i is calculated from the previous x_i's.

In many information retrieval systems, the rank of the document *D* with respect to the query *Q* is calculated as

$$S(D, Q) = \sum_{x\in Q} w_{xQ}.w_{xD}.w_{xC} \quad (5)$$

where w_{xQ} is the weight of term *x* in query *Q*, w_{xD} is the weight of term *x* in document D and w_{xC} is the weight of term *x* in corpus C that usually the IDF relationship is used to calculate it. In this paper, the focus has been on w_{xC} and its value has been calculated by using linear time series.

According to the hypothesis offered in Refrence [5], the frequencies of terms that have less importance in the information retrieval (like term 'the') is dependent on a special time series and this dependency decreases as the importance of the term increases. Fig. 1 shows the frequency of two terms 'the' and 'schengen' from LA Times corpus in 100 consecutive weeks. As can be seen, the frequency diagram for 'the' is strongly linear. However, the frequency of 'schengen' does not follow a linear relationship. According to the hypothesis of this paper, this point could be used for term weighting. To do this, the weight of each term is calculated by

$$\text{Weight (x)} = \sqrt{\sum_{i=1}^{n}(\hat{x}_i - x_i)} \quad (6)$$

where x_i is the frequency of *x* at time *i* and $\hat{x}_i$ is the the expected frequency of term *x* at time *i*. If we use linear time series to calculate $\hat{x}_i$, it is possible to evaluate the offered hypothesis in the paper; because as the dependency level of the term frequency is less, the value of weight *x* will be lessened and according to the hypothesis, the term will have lower weight. In contrast, if the frequency of the term does not follow linear time series, it will have a higher weight. For example, the weight of the term 'the' is less than 'schengen'.

To calculate $\hat{x}_i$, there are several methods that one of them is called Moving Average which each value of $\hat{x}_i$ is calculated as

$$\hat{x}_i(x_{i-1} + x_{i-2} + \cdots + x_{i-p})/p \quad (7)$$

In other words, the expected value of $\hat{x}_i$ is obtained from the average of *p* previous observations where *p* is a configurable parameter. Empirical experiments in the paper on four different data sets have shown that term weighting based on linear time series have better results considering the relation of extracting documents with the queries.

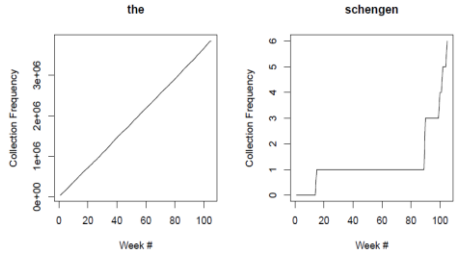

Fig. 1. Time Series of the two terms 'the' and 'schengen' from LA Times TREC Corpus [5].

C. Term Weighting Based on Global Analysis and Burst Revision History

Reference [3] focuses on the document changes during the time. In the most of term weighting methods, documents have been considered statically and each change in a document is considered as a new document. However, this change will cause creating a revision of the same document. Reference [3] solves this problem by using two solutions. First, the document changes have been considered over the whole period which has been discussed under the title of *global revision history analysis*. Second, it is possible that temporal changes of a document cause changing a term weight. For example, if a document is edited several times in a one day interval and a term is used frequently, but in the next days the term is deleted; it will lose its importance. Therefore, another analysis method has been provided that this issue has been considered in which is named *revision history burst analysis*.

To global analysis of revision history, a new concept for the term frequency has been introduced called *global term frequency* which for each term *t* in document *d* (the document that is revised) is calculated by

$$TF_{global}(t,d) = \sum_{j=1}^{n} \frac{c(t,v_j)}{j^{\alpha}} \quad (8)$$

wherer *j* is the number of document revisions, which has been numbered from 1 to *n* chronologically and $c(t,v_j)$ is the raw frequency of term *t* in the revision v_j. j^{α} is called *decay factor* where α controls the speed of the decay. If $\alpha>0$, the weight of a term will decrease in later revisions. In contrast, if $\alpha<0$, the weight of a term will decrease in later revisions (in their experiments, the optimal value for α was 1.1).

To revision history burst analysis, *term burst factor* concept has been provided where for each term *t* in document *d* that has burst time intervals like $\{b_1, b_2, ..., b_m\}$ is calculated by

$$TF_{burst}(t,d) = \sum_{j=1}^{m} \sum_{k=b_j}^{n} \frac{c(t,v_k)}{(k-b_j+1)^{\beta}} \quad (9)$$

where $c(t,v_k)$ is the raw frequency of term *t* in the revision v_k and $(k-b_j+1)^{\beta}$ is the decay factor. Therefore, when a burst b_j happens ($k=b_j$), this factor will be equal to 1 and the impact of this burst will decrease steadily; because the decay factor increases by the growth of k (in their empirical experiments the value of β was 1.1).

To apply above relations in the existing models, (10) has been offered to calculate the weight of each term *t* in document *d*:

$$\mathrm{TF_{RHA}}(\mathrm{t,d}) = \lambda_1 \mathrm{TF_{global}}(\mathrm{t,d}) + \lambda_2 \mathrm{TF_{burst}}(\mathrm{t,d}) + \lambda_3 \mathrm{TF}(\mathrm{t,d}) \quad (10)$$

where $\mathrm{TF_{global}}(\mathrm{t,d})$ and $\mathrm{TF_{burst}}(\mathrm{t,d})$ are the measures defined in the paper and $\mathrm{TF}(\mathrm{t,d})$ is the measure defined in the existing model. λ_1, λ_2, and λ_3 are configurable parameters so that $\lambda_1 + \lambda_2 + \lambda_3 = 1$. Experimental results in the paper on two different dataset have been shown that revision history analysis provides consistent improvement over the similarity of extracted documents to the queries.

D. Term Weighting Based on Document Revision Histroy

Reference [8] considers document revisions and different measurements. These measurements include *Revision Frequency*, *Revision Term Frequency*, *Relative Term Frequency*, *Revision Span*, and *Revision Term Frequency Span* which have been discussed in the following:

- **Revision Frequency (rf):** This measure is calculated by

$$\mathrm{rf_{t,d}} = \frac{|R_{t,d}|}{|R_d|} \quad (11)$$

where $R_{t,d}$ is the number of revisions that include term *t* and R_d is the total number of revisions.

- **Relative Term Frequency (rel_tf):** This measure is calculated by

$$\mathrm{rel\_tf_{t,d}} = \frac{\mathrm{tf_{t,d}}}{\sum_{\mathrm{v}\in\mathrm{d}} \mathrm{tf_{v,d}}} \quad (12)$$

In this measure the impact of the other terms of document *d* has been considered, too.

- **Revision Span (rs)**: As a revision time span has not been considered in the above relations, it has been solved in

$$rs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} (ts(r_{i+1}) - ts(r_i))}{ts(r_n) - ts(r_1)} \quad (13)$$

where $ts(r_i)$ is the i_{th} revision time of document *d*.

- **Revision Term Frequency Span**: This measurement is calculated as

$$rtfs_{t,d} = \frac{\sum_{r_i \in R_{t,d}} (rel\_tf_{t,r_i} * (ts(r_{i+1}) - ts(r_i)))}{ts(r_n) - ts(r_1)} \quad (14)$$

Experimental evaluations on three different datasets have shown that the above measures have a better performance compared to the usual term frequency formula.

III. The Proposed Method to Improve Term Weighting Efficiency

In all presented methods in section II, to measure the weight of terms, the whole revision(s) of a document should be considered that could be time consuming. Refrence [3], discusses that the similarity of two consecutive revisions increases during time using its

proposed method. For example, the similarity of two last revisions of a document is more than the similarity of two first revisions. According to our hypothesis, it is possible to use the aforementioned fact to improve the speed of term weighting. To do this, we used only a special part of revisions in weighting. To choose this special part, we tried to choose more instances from the initial revisions and less instances from the last ones; because there is more similarity between the last revisions. Therefore, to reduce the computation time, it is possible to choose fewer instances from last revisions.

Fig. 2 proposes our algorithm to choose this special part. In this algorithm, p sets the amount of jumps over the revisions. For example, if $p=50$, the first fifty revisions of revisions will be selected; because the value of $\left\lfloor sqrt(\frac{i}{p}) \right\rfloor$ will be equal to zero. But when the value of i is 50, the new value of i will be equal to 52 that causes the 51th revision will not be selected. In the proposed algorithm, the gap between two selected revisions grows as the value of i increases. This causes more initial revisions are selected and the number of the selected revisions decreases as we move toward the last ones.

1. Set $i = 1$ and $selected = \{\}$
2. Add $revision[i]$ to $selected$
3. while $(i < number\_of\_revisions)$
4. {
5. $i = i + 1 + \left\lfloor sqrt(\frac{i}{p}) \right\rfloor$
6. Add *revision[i]* to *selected*
7. }

Fig2. The algorithm of selecting revisions.

IV. Empirical Evaluation of the Proposed Method

In this section, we present the empirical results of the proposed method on three different datasets. To do this, the results of the introduced measures in subsection 2.D on set of selected revisions and the whole set of revisions are compared together.

A. Datasets

To evaluate the introduced measures, we chose three independent data sets from the English version of Wikipedia. The first set includes random documents from the featured documents of Wikipedia that usually a lot of revisions are performed on. The second set includes documents that have been obtained by using the Random Feature" of Wikipedia. The third set includes a random subset of the document set of Wikipedia which have been tagged by different users in a popular tagging website. Refrence [11] introduces this set and include more than 20000 independent Wikipedia documents. The information about these three sets has been presented in Table 1.

Table 1. Statistical Information About the Each Three Datasets.

Set	The number of Documents	The Average Number of Revisions
Featured	20	520
Random	20	185
Social	20	661

B. Evaluation of Measures on the Featured Dataset

To evaluate the provided measures on the featured dataset, the abstract section of each document on Wikipedia has been used. According to the guide of producing documents on Wikipedia, the abstract section should include the title and abstract of each document's content. For this purpose, the terms with the highest weights were selected and the cosine similarity of these terms with the abstract section has been used as a criterion to evaluate the measures. To calculate the cosine similarity, the frequency vector of terms was made that for terms with the highest weight, term frequency in the whole history and for the abstract section, term frequency in the abstract section of the last revision have been calculated. Fig. 3 And Fig. 4 present the evaluation results of the measures on the whole revision history and the selected revision history, according to the provided algorithm in section III, respectively (the value of p is 50). In these diagrams, the x-axis shows number of terms with the highest weight and the Y-axis shows the cosine similarity of terms with the terms of the abstract section of the document. As can been seen in the both diagrams, the introduced measures have better efficiency compared to the measure of term frequency.

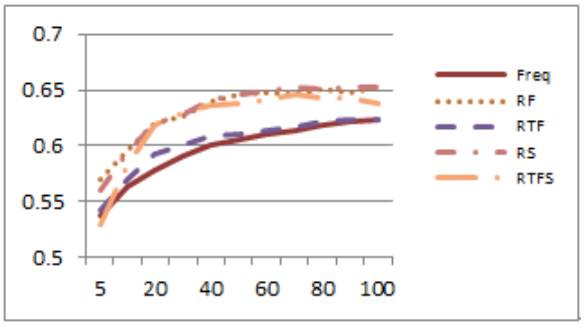

Fig. 3. The average percentage of similarity between the high ranked terms and the terms in the abstract section of the document on the whole history of revisions.

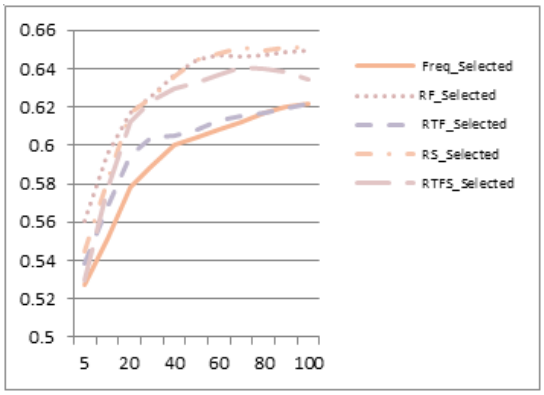

Fig. 4. The average percentage of similarity between high ranked terms and the terms of abstract section on the selected history of revisions.

Fig. 5, shows the results of each measure on the selected revision set and the term frequency on the whole history of revisions. As can be seen, although some part of the revision history, has been selected only, *rf*, *rtf*, and *rs* measures have higher efficiency in the selected revision history compared to the term frequency measure. However, this difference has not been too much and of course has been expected. The average number of selected revisions has been 278 and the average number of the whole revisions has been 520.

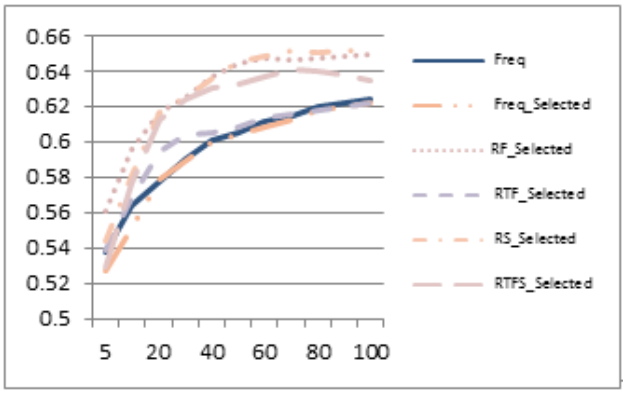

Fig. 5. The average percentage of similarity according to the introduced measures on the selected revisions and the term frequency measure on the whole revision history.

C. Evaluation of Measures on Random Dataset

To evaluate provided measures on the random dataset, the abstract section of each document has been used, too. Fig. 6 presents the evaluation results. It confirms that *rf*, *rtf*, and *rs* measures have better efficiency in the selected revision history compared to the term frequency measure in the whole revision history. Here, the average number of selected revisions has been 134 and the average number of the whole revision history has been 185.

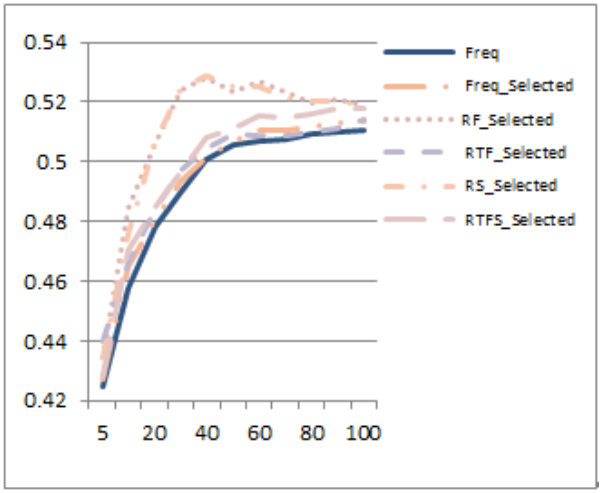

Fig. 6. The average percentage of similarity for the random dataset.

D. Evaluation of Measures on Social Dataset

To evaluate provided measures on a social dataset, tags which users on the *delicious website* have used to specify pages, have been employed. To do this, terms which have had the highest weights according to each measure have been selected and their cosine similarity with the chosen tags by users has been considered as a criterion to evaluate measures. Fig. 7 shows the results of this evaluation. It confirms that *rf*, *rtf*, and *rs* measures have better efficiency in the selected revision history compared to the term frequency measure in the whole revision history. Here, the average number of selected revisions has been 317 and the average number of the whole revision history has been 661. In other words, less than half of terms have been investigated.

The reason of similarity reduction in the cases which more than 30 terms are selected is the limitation of the *delicious* website in providing the number of tags. On this website, the maximum number of the tags for a document is 30. Therefore, in case of choosing more than 30 terms of those terms with the highest weights, the level of cosine similarity will be decreased.

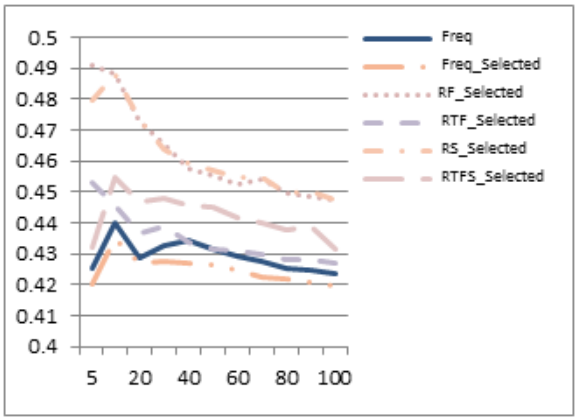

Fig. 7. The average percentage of similarity for social dataset.

V. Conclusion and Future Works

In this paper, to improve the efficiency of the term weighting method, a special part of the revision history of documents was considered. The experimental results on the chosen datasets show that using the proposed method decreases the analysis time of the document set. Nevertheless, the quality of information retrieval has been almost kept constant.

The results of this research can be used in search engines, too. As search engines refer to different web pages on websites periodically and update their information, the proposed method can decrease unnecessary processes for document ranking.

In the proposed method, the criterion to select the revisions is the revision number. Therefore, one possible future work could be considering the lifetime of each revision to choose the history of revisions and evaluating its impact on the method efficiency and the quality of information retrieval.

References

[1] E. Adar, J. Teevan, and S. T. Dumais, "Resonance on the Web: Web Dynamics and Revisitation Patterns," in

Proceedings of CHI 2009, 2009, "doi: 10.1145/1871437.1871519".

[2] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas, "The Web Changes Everything: Understanding the Dynamics of Web Content," in *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009, pp. 282–291, "doi: 10.1145/1498759.1498837".

[3] A. Aji, Y. Wang, E. Agichtein, and E. Gabrilovich, "Using the Past to Score the Present: Extending Term Weighting Models Through Revision History Analysis.," in *CIKM*, 2010, pp. 629–638, "doi: 10.1145/1871437.1871519".

[4] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, "Survey of Temporal Information Retrieval and Related Applications," *ACM Comput. Surv.*, vol. 47, no. 2, pp. 15:1–15:41, 2014, "doi: 10.1145/2619088".

[5] M. Efron, "Linear Time Series Models for Term Weighting in Information Retrieval.," *JASIST*, vol. 61, no. 7, pp. 1299–1312, 2010, "doi: 10.1002/asi.21315 ".

[6] J. L. Elsas and S. T. Dumais, "Leveraging Temporal Dynamics of Dcument Content in Relevance Ranking.," in *WSDM*, 2010, pp. 1–10, "doi: 10.1145/1718487.1718489".

[7] N. Kanhabua, "Time-aware Approaches to InformationRetrieval," *SIGIR Forum*, vol. 46, no. 1, p. 85, 2012, "doi: 10.1145/2215676.2215691.

[8] Nunes, C. Ribeiro, and G. David, "Term Weighting Based on Document Revision History.," *JASIST*, vol. 62, no. 12, pp. 2471–2478, 2011, "doi: 10.1002/asi.21597".

[9] Nunes, C. Ribeiro, and G. David, "Term Frequency Dynamics in Collaborative Articles," in *Proceedings of the 10th ACM Symposium on Document Engineering*, 2010, pp. 267–270, "doi: 10.1145/1860559.1860620".

[10] K. Radinsky, F. Diaz, S. Dumais, M. Shokouhi, A. Dong, and Y. Chang, "Temporal Web Dynamics and Its Application to Information Retrieval," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, 2013, pp. 781–782.

[11] A. Zubiaga, "Enhancing Navigation on Wikipedia with Social Tags," *CoRR*, vol. abs/1202.5, 2012.

[12] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," *Inf. Process. Manag. an Int. J.*, vol. 24, no. 5, pp. 513–523, 1988, "doi: 10.1016/0306-4573(88)90021-0".

Authors' Profiles

Mehdi Jabalameli received the B.Sc. degree in Computer Engineering, Kharazmi University, Tehran, Iran, an MSc in Computer Engineering, Sharif University of Technology, Tehran, Iran. Currently Mehdi Jabalameli is a PhD candidate in the Department of Computer Engineering, University of Isfahan, Isfahan, Iran. His research interest includes Information Security, Information Retrieval, Semantic Web, Linked Data.

Ala Arman received his B.S.c degree in Computer Engineering at the Azad University, Najafabad Branch, Esfahan, Iran and his Msc in Software Engineering of Distributed Systems at the KTH (Royal Institute of Technology), Stockholm, Sweden. He is a PhD candidate in the University of Milan. His research interests include distributed systems, cloud computing and software engineering.

Mohammad Ali Nematbakhsh is an associate professor of computer engineering Department at the University of Isfahan. He received his B.Sc. in Electrical Engineering from Louisiana Tech University in 1981 and his M.Sc. and PhD degrees in Electrical and Computer Engineering from the University of Arizona in 1983 and 1987, respectively. He worked for Micro Advanced Co. and Toshiba Corporation in USA and Japan for many years before joining The University of Isfahan. He has published more than 100 research papers, three U.S. registered patents and a database book that is widely used in universities. His main research interests include semantic web and database systems.

How to cite this paper: Mehdi Jabalameli, Ala Arman, Mohammadali Nematbakhsh,"Improving the Efficiency of Term Weighting in Set of Dynamic Documents", IJMECS, vol.7, no.2, pp.42-47, 2015.DOI: 10.5815/ijmecs.2015.02.06

